CAN WE PREDICT THE SUCCESS OF A MOVIE?

The 🤚🤚 Team

ANSWERING THE WHY

Who is the YAY! Team?

Ally Qi
Anastasia Valdes
Grace Huefner
Kellie Cox
Masita Mohamad

Where did the data come from?

We found a 5000-movie database on Kaggle that queries fields from The Movie Database (a site like IMDB)

Source: https://www.kaggle.com/tmdb/tmdb-movie-metadata

What kinds of questions did we ask?

Can we predict the success of a movie based on:

Release date
Runtime
Geography
Genre
Title Length

How did we measure success?

Vote Rating: the average rating that the movie received on the site

Adjusted Revenue: revenue adjusted for inflation in 2020

CLEANING AND EXPLORING THE DATA

- ✅ Find the dataset of interest
- ✅ Check data types
- ✅ Check data sample size
- ✅ Check counts for each variable
- ✅ Remove unused variables
- ✅ Check summary statistics
- ✅ Check missing values
  - Imputation for missing values
  - Replaced with mean
- ✅ Confirm no duplicates
- ✅ Set unique identifier to "ID"
  - Final sample size: 4800

Created a "decade" variable

Created a "release year" variable
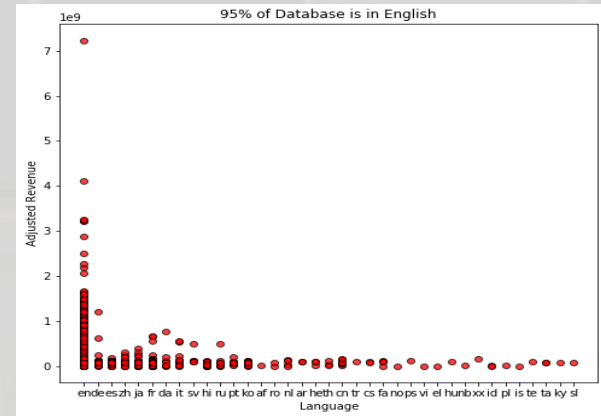
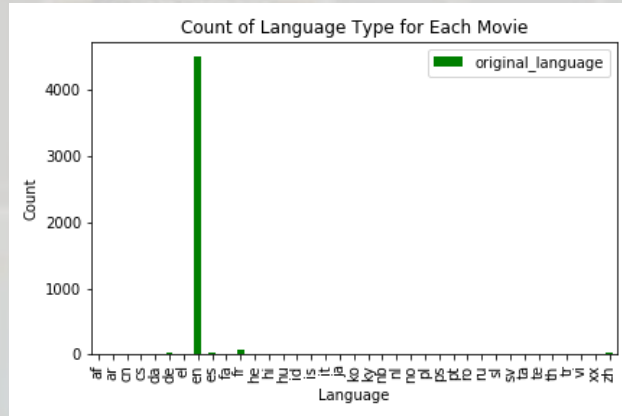**Adjusted inflation for budget and revenue***
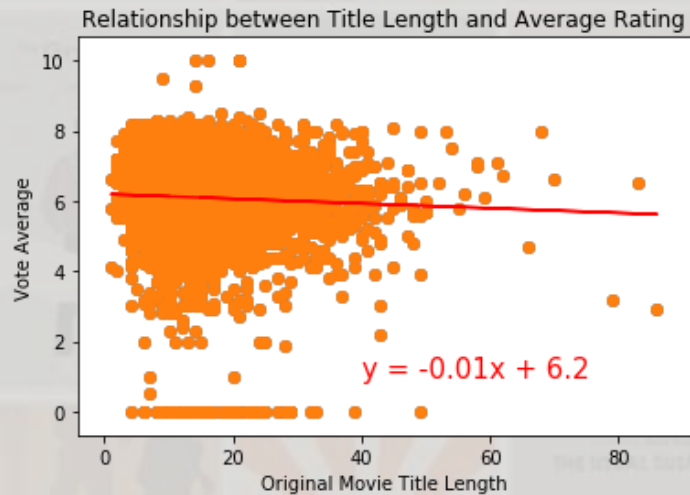- Import cpi

Extract genres to its own column*
- True or False
- 20 genres found!

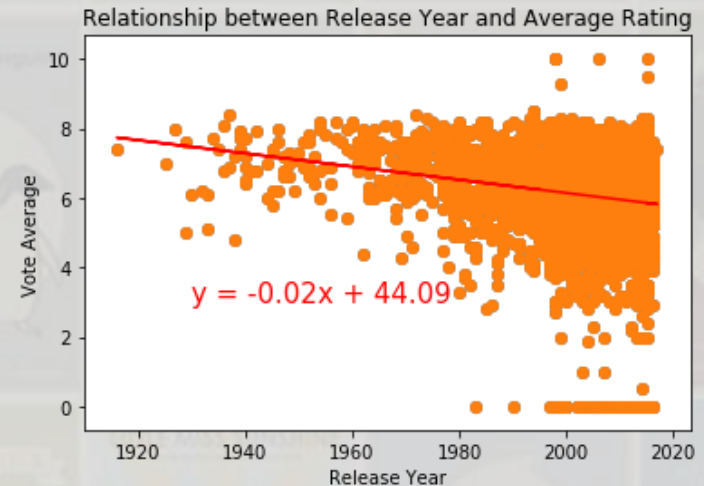# SOME QUESTIONS WERE D-U-Ds…

What movies are the most popular by geography?

What movies are most popular by title length?

What movies are most popular by release year?
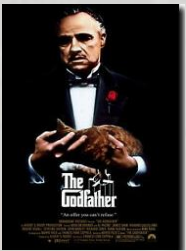
SOME QUESTIONS JUST NEEDED MORE...

Vote Average

Success Measurement

Revenue

What is the trend between decade and movie genre? Have we seen a genre resurgence?

Top 5 Movies of the 1970s – By Vote Average

1. The Godfather
Vote Average: 8.4
Movie Genres: Crime, Drama

2. The Godfather: Part II
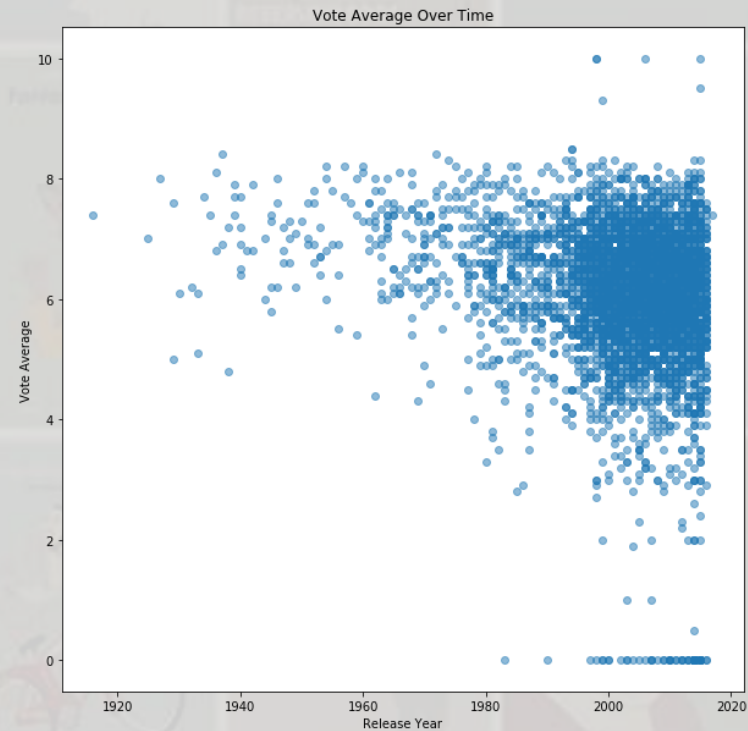Vote Average: 8.3
Movie Genres: Crime, Drama

3. One Flew Over the Cuckoo's Nest
Vote Average: 8.2
Movie Genres: Drama

4. Star Wars
Vote Average: 8.1
Movie Genres: Action, Adventure, Science Fiction

5. Apocalypse Now
Vote Average: 8.0
Movie Genres: Drama, War

Vote Average Over Time

SOME QUESTIONS JUST NEEDED MORE...

Success Measurement
- Vote Average
- Revenue

What is the trend between decade and movie genre? Have we seen a genre resurgence?

Top 5 Movies of the 1970s – By Revenue

1. Star Wars
Revenue: $3.23B
Movie Genres: Action, Adventure, Science Fiction
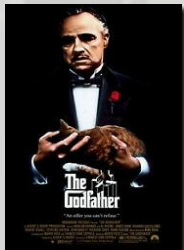

2. The Exorcist
Revenue: $2.49B
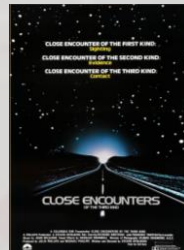Movie Genres: Drama, Horror, Thriller

3. Jaws
Revenue: $2.2B
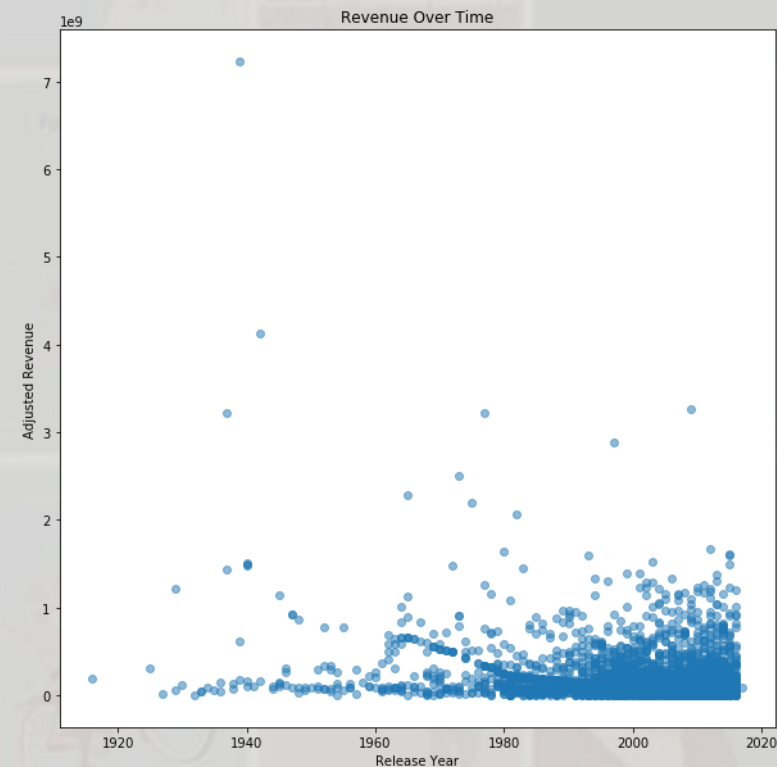Movie Genres: Adventure, Horror, Thriller

4. The Godfather
Revenue: $1.5B
Movie Genres: Crime, Drama

5. Close Encounters of the Third Kind
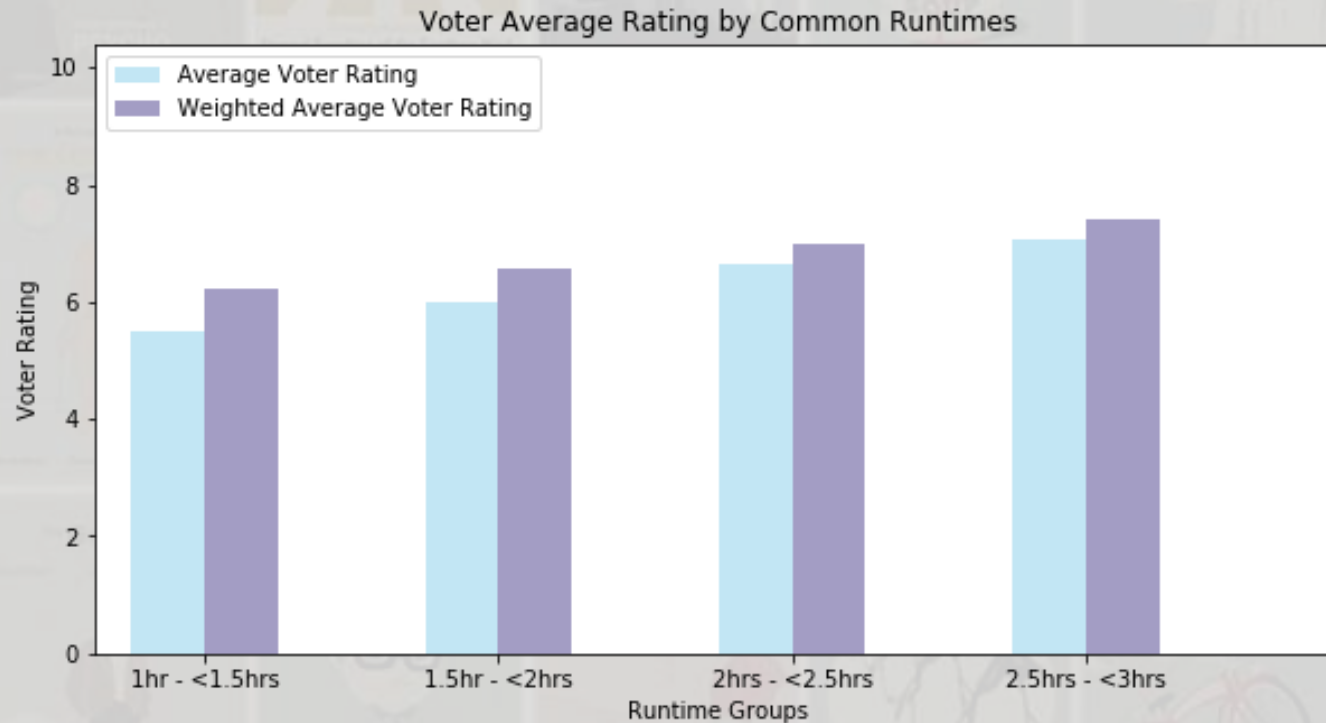Revenue: $1.25B
Movie Genres: Drama, Science Fiction

Revenue Over Time

Image Credit: Etsy

AND SOME QUESTIONS MADE SENSE!

Vote Average

Success Measurement

Revenue

What is the relationship between Runtime and Vote Average?

Voter Average Rating by Common Runtimes

- The average runtime of movies on the list was 108 minutes or 1hr 48 mins.
- Those movies between 1 hour and 3 hours received the most votes, and therefore were deemed the most common.
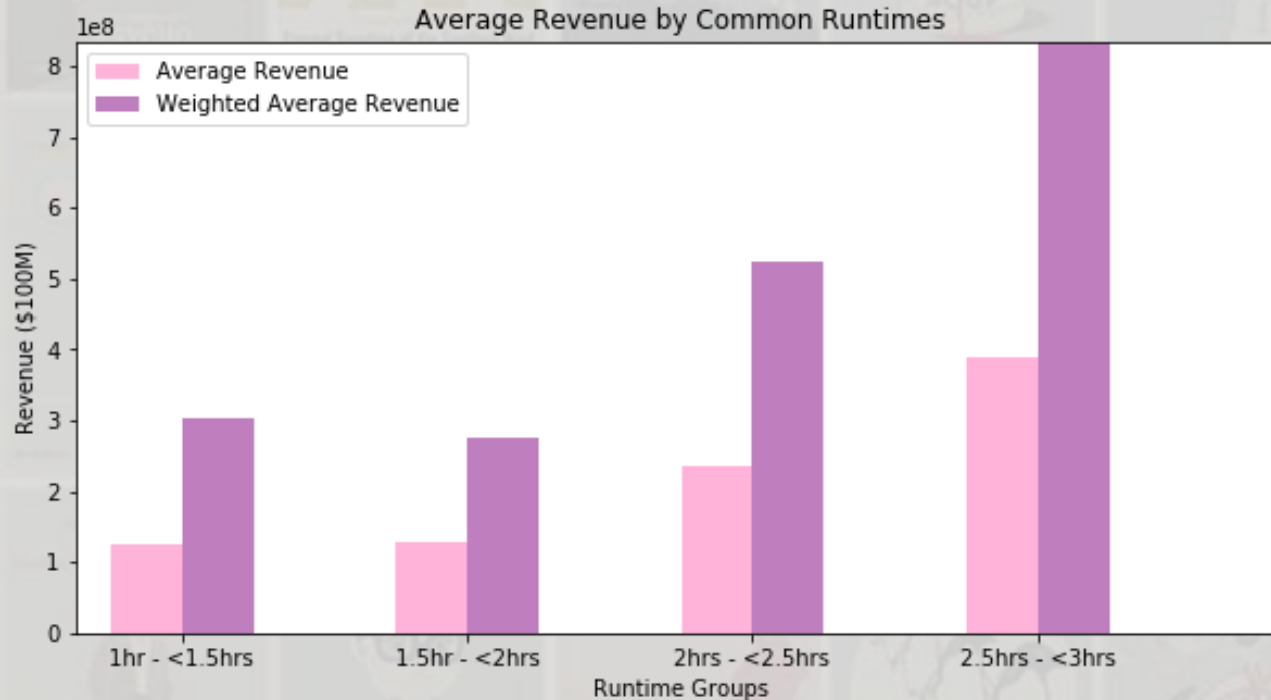
Image Credit: Etsy

AND SOME QUESTIONS MADE SENSE!

Success Measurement
- Vote Average
- Revenue

What is the relationship between Runtime and Revenue?

Average Revenue by Common Runtimes

- Those movies between 1 hour and 3 hours received the most votes, and therefore were deemed the most common.
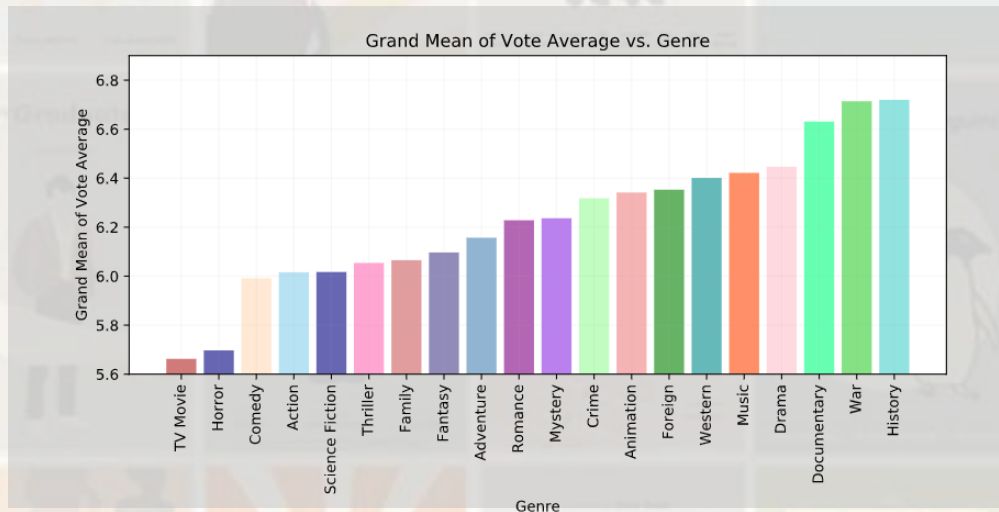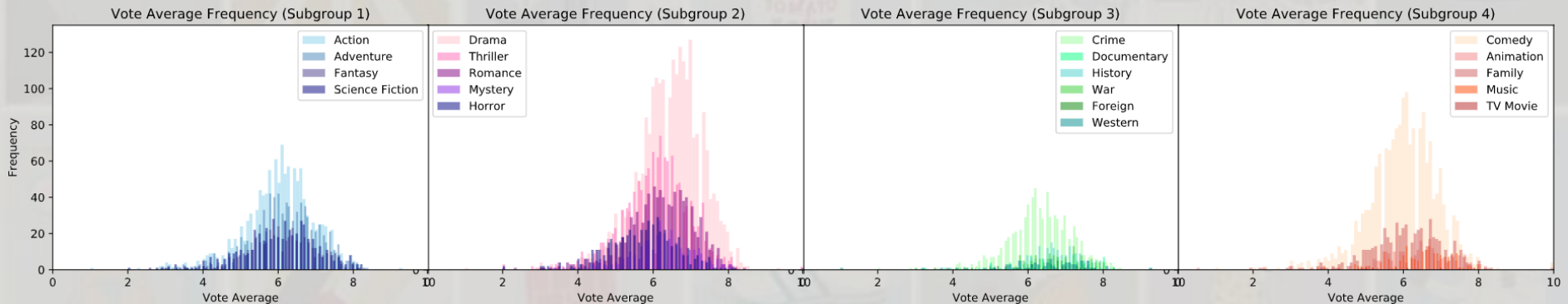- Longer movies made more revenue in the box office

AND SOME QUESTIONS MADE SENSE!

Success Measurement → Vote Average / Revenue

What is the relationship between Genre and Vote Average?

Vote Average Frequency (Subgroup 1): Action, Adventure, Fantasy, Science Fiction

Vote Average Frequency (Subgroup 2): Drama, Thriller, Romance, Mystery, Horror

Vote Average Frequency (Subgroup 3): Crime, Documentary, History, War, Foreign, Western

Vote Average Frequency (Subgroup 4): Comedy, Animation, Family, Music, TV Movie

Grand Mean of Vote Average vs. Genre

- Drama received the most amount of votes. However, it is also the broadest definition of all genres
- Documentary, War, History received less but more favorable votes. Audiences seem to enjoy fact-based movies
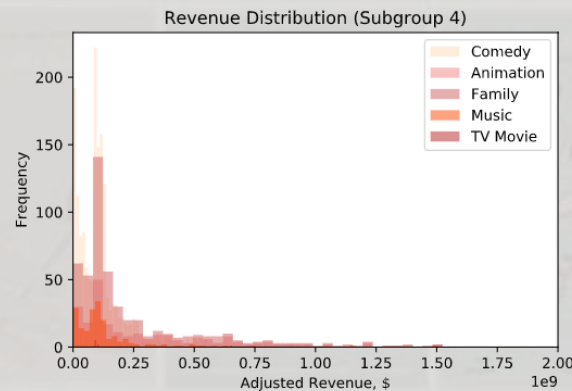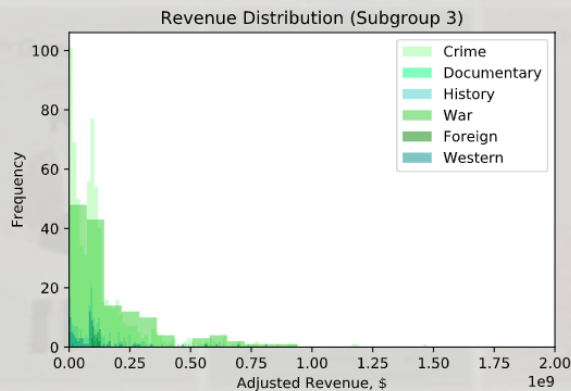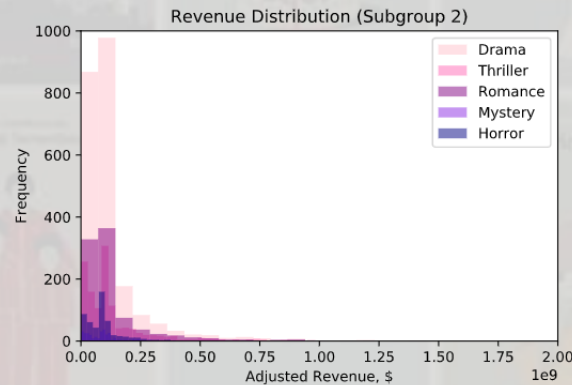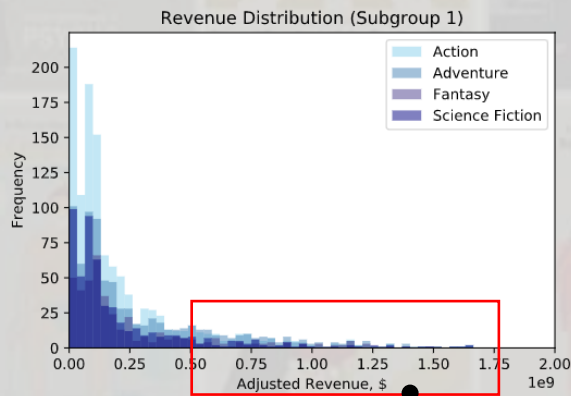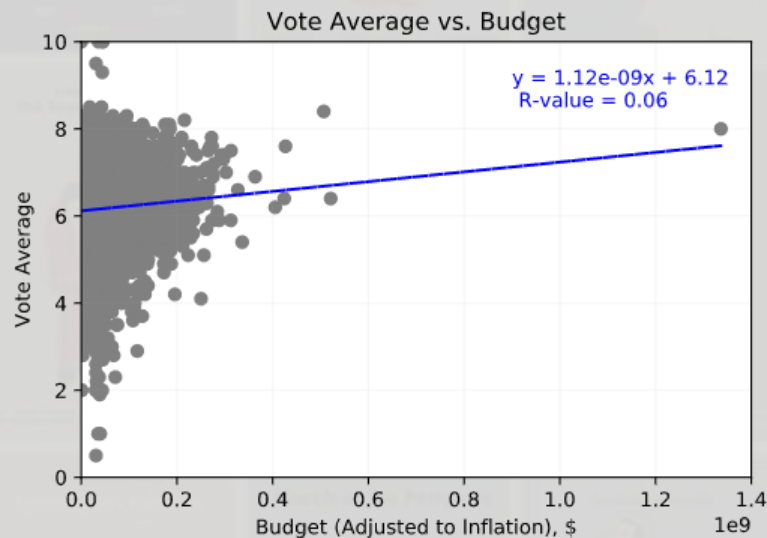
AND SOME QUESTIONS MADE SENSE!

Success Measurement
- Vote Average
- Revenue

What is the relationship between Genre and Revenue?

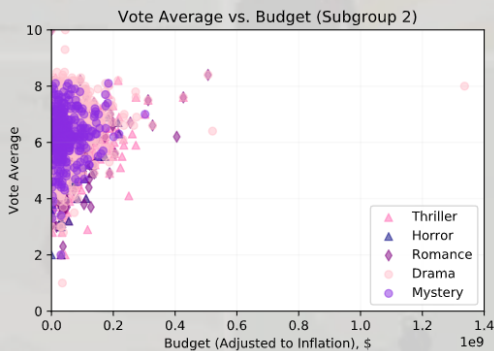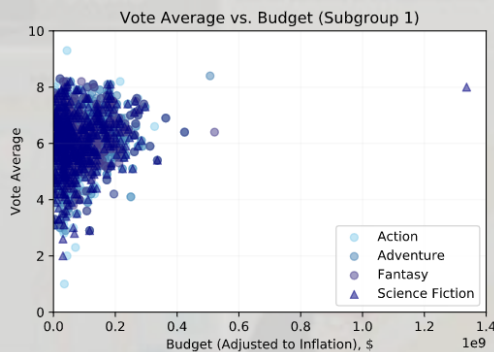Subgroup 1 brings in most high revenue range (>$500M) vs other genre
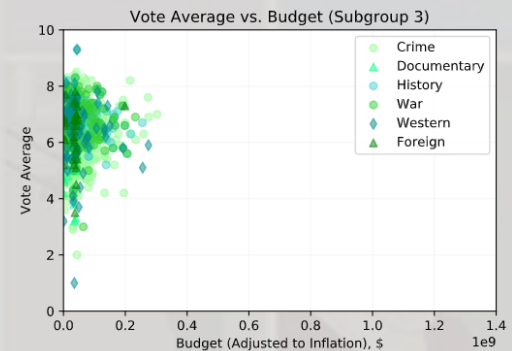
AND SOME QUESTIONS MADE SENSE!

What is the relationship between Production Budget and Vote Average?

Vote Average vs. Budget (Subgroup 1) — Action, Adventure, Fantasy, Science Fiction

Vote Average vs. Budget (Subgroup 2) — Thriller, Horror, Romance, Drama, Mystery

Vote Average vs. Budget

$$y = 1.12\text{e-}09x + 6.12$$
$$R\text{-value} = 0.06$$

Vote Average vs. Budget (Subgroup 4) — Comedy, Animation, Family, Music, TV Movie

Vote Average vs. Budget (Subgroup 3) — Crime, Documentary, History, War, Western, Foreign

- There is a slight positive relationship between production cost and viewers' ratings.
- Subgroup 1 (Action, Adventure, Fantasy & Science Fiction) is more expensive to produce possibly due to special sets, props, visual and special effects etc, however, this seems to not have much bearings on the viewers' votes

Image Credit: Etsy

AND SOME QUESTIONS MADE SENSE!

Success Measurement → Vote Average, Revenue
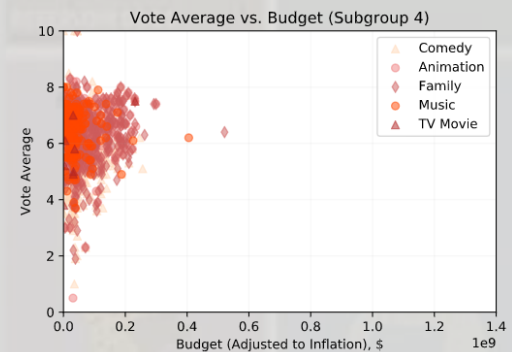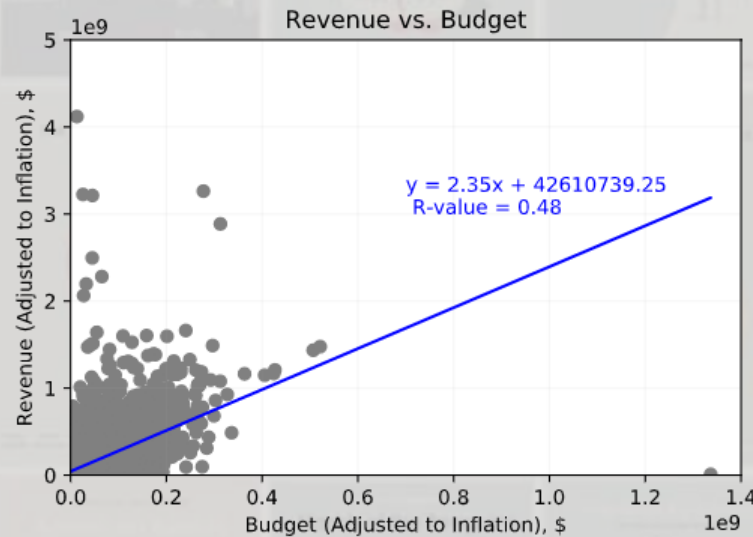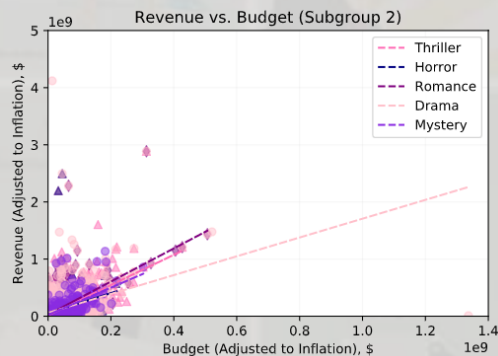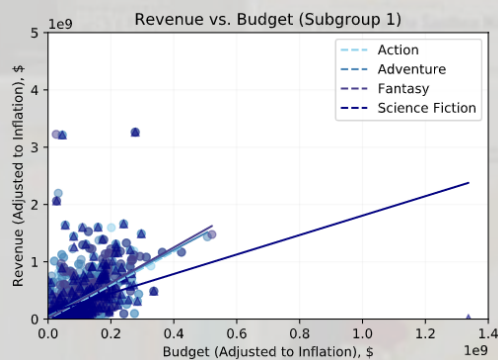
What is the relationship between Production Budget and Revenue?

Revenue vs. Budget (Subgroup 1) — Action, Adventure, Fantasy, Science Fiction

Revenue vs. Budget (Subgroup 2) — Thriller, Horror, Romance, Drama, Mystery

Revenue vs. Budget
$y = 2.35x + 42610739.25$
R-value = 0.48

Revenue vs. Budget (Subgroup 4) — Comedy, Animation, Family, Music, TV Movie

Revenue vs. Budget (Subgroup 3) — Crime, Documentary, History, War, Western, Foreign

- There is a strong relationship between production cost and production revenue
- Bigger film production budget seems to result in higher revenue
- Western movies produces the lowest Return of Investment

Image Credit: Etsy
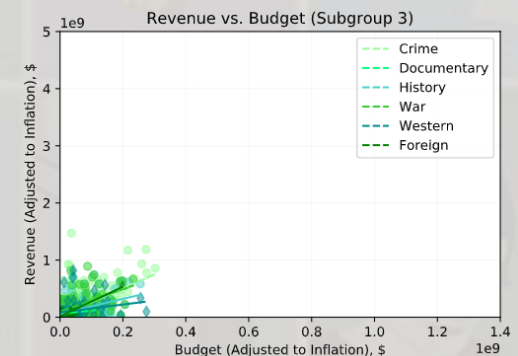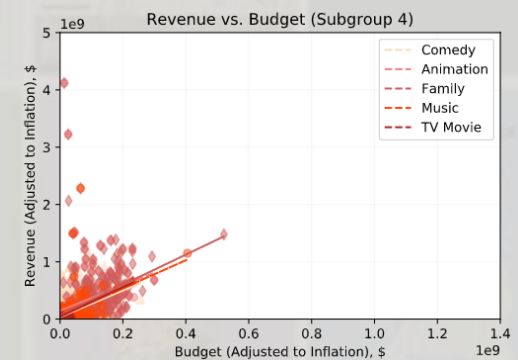
SO... CAN WE PREDICT THE SUCCESS?

Yes!

By using very specific parameters set for this data we can.

If you're going by vote average

Make a drama or fact-based movie, that is between 1.5 – 2.5 hours long and has a budget above $200 M.

If you're going by box office revenue

Make an action, adventure, or fantasy that is between 1.5 – 2.5 hours long and has a budget above $200 M.

Ultimately, we found the data set was not ideal for what we were looking to answer. Sites like IMDB are likely to have more detailed results and findings that we could use to help support our hypothesis

Image Credit: Etsy